Estimation de quantiles extrêmes et probabilites d'evènements rares d'un processus stochastique

Gilles Durrieu, Ion Grama, Quang-Khoai Pham, Tricot Jean-Marie

Estimation de quantiles extrêmes et probabilités d'évènements rares d'un processus stochastique

Gilles Durrieu, Ion Grama, Quang Khoai Pham et Jean-Marie Tricot

Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205
Campus de Tohannic, 56017 Vannes.
{gilles.durrieu, ion.grama, quang-khoai.pham,jean-marie.tricot}@univ-ubs.fr

Résumé. Nous considérons un processus stochastique à temps continu $X(t)$ à incréments indépendants de distribution F_t. Nous proposons un estimateur adaptatif non paramétrique de quantiles d'ordre élevés. L'idée de notre approche consiste à ajuster la queue de la distribution F_t, avec une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimée en utilisant un estimateur à noyau de taille de fenêtre h basé sur les observations plus grandes que τ. Sous certaines hypothèses de régularité, nous montrons que l'estimateur adaptatif proposé de $\theta_{t,\tau}$ est consistant et nous donnons sa vitesse de convergence. Nous proposons également une procédure de tests séquentiels pour déterminer le seuil τ et le paramètre h. Enfin, nous étudions les propriétés de cette méthode sur des simulations et sur des données réelles dans le but d'estimer des changements globaux (pollution, changement de température) et ainsi d'aider à la surveillance de systèmes aquatiques.

Mots-clés. Statistique environnementale, valeurs extrêmes, Données à Haute fréquence, estimateur non paramétrique à noyau.

Abstract. Consider a continuous time process $X(t)$ with independent increments distributed according a distribution function F_t. We propose a nonparametric adaptive estimator of the high quantiles. The idea of our approach is to adjust the tail of the distribution function F_t with a Pareto distribution with parameter $\theta_{t,\tau}$ starting from a threshold τ. The parameter $\theta_{t,\tau}$ is estimated using a kernel estimator of bandwidth h based on the observations larger than τ. Under some regularity assumptions, we prove that the proposed adaptive estimator of $\theta_{t,\tau}$ is consistent and we determine its rate of convergence. We also propose a sequential tests based procedure for the authomatic choice of the threshold τ and the bandwidth h. Finally, we study the properties of this method by simulation and on real data set to estimate global changes (pollution, temperature change) and so to help in the survey of aquatic systems.

Keywords. Environmental statistics, extreme values, high frequency data, non parametric kernel estimator.

1 Modèle et estimateurs

Nous considérons un processus stochastique à temps continu $X(t)$, $t \in [0,T]$, à incréments indépendants de distribution F_t définie sur $[x_0, +\infty[$ avec $x_0 \geq 0$. On suppose que F_t appartient à l'ensemble $\mathcal{F}$ des fonctions de répartition strictement croissantes admettant une densité continue par rapport à la mesure de Lebesgue μ. Dans ce papier, nous proposons une méthode d'estimation adaptative de la queue de distribution de F_t et des quantiles extrêmes, basée sur une procédure séquentielle de tests d'adéquations. L'idée de la méthode est de déterminer de manière adaptative un seuil τ en ajustant sur $[\tau, +\infty[$ une famille paramétrique de modèles $\{\mathcal{P}_\theta; \theta > 0\}$. On choisit ici pour $\mathcal{P}_\theta$ une distribution de Pareto définie par

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}}, x \in [\tau, +\infty[,$$

où le paramètre $\theta > 0$ et $\tau \geq x_0$ est la valeur inconnue du seuil. Par conséquent, nous proposons un modèle semi-paramétrique défini par

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x < \tau \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x \geq \tau. \end{cases} \quad (1)$$

Soit $\mathcal{K}(P,Q) = \int \log \frac{dP}{dQ}\, dP$ la divergence de Kullback-Leibler entre deux mesures équivalentes P et Q. Pour chaque $t \in [0,T]$ et $\tau \geq x_0$, le minimum de la divergence de Kullback-Leibler entre F_t et le modèle $F_{t,\tau,\theta}$ est atteint pour

$$\theta_{t,\tau} = \arg\min_{\theta \in \Theta} \mathcal{K}(F_{t,\tau}, G_{\tau,\theta}) = \int_\tau^\infty \log \frac{x}{\tau} \frac{F_t(dx)}{1 - F_t(\tau)}.$$

Pour chaque t fixé dans $[0,T]$ et pour $\tau \geq x_0$, nous construisons un estimateur non paramétrique à noyau K de taille de fenêtre h du paramètre fonctionnel $t \to \theta_{t,\tau}$. Nous estimons dans un premier temps la fonction $\theta_{t,\tau}$ au point t en utilisant un estimateur à noyau d'une taille de fenêtre h et dans un second temps nous donnons une procédure de sélection du seuil τ. En maximisant la quasi-log vraisemblance pondérée par rapport à θ, nous obtenons l'estimateur

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{X_{t_i} > \tau} W_{t,h}(t_i) \log\left(\frac{X_{t_i}}{\tau}\right), \quad (2)$$

où $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$ avec K un noyau Gaussien et $\widehat{n}_{t,h,\tau} = \sum_{i=1}^n W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}}$. L'estimateur semi-paramétrique de la fonction de répartition F_t est donné par

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x), & x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right)\left(\frac{x}{\tau}\right)^{-\frac{1}{\widehat{\theta}_{t,h,\tau}}}, & x > \tau. \end{cases}$$

où

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i)\, 1_{\{X_{t_i} \leq x\}}$$

est la fonction de répartition empirique pondérée. L'estimateur semi-paramétrique du quantile d'ordre p est donné par

$$\widehat{q}_p(t) = \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{pour} \quad p < \widehat{p}_0, \\ \tau \left(\frac{1-\widehat{p}_0}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{sinon,} \end{cases} \tag{3}$$

avec $\widehat{p}_0 = \widehat{F}_{t,h}(\tau)$.

L'estimateur $\widehat{\theta}_{t,h,\tau}$ est très sensible aux choix du seuil τ. La difficulté est de choisir τ assez petit de façon à ce que l'estimateur de la fonction de répartition empirique pondérée dans le modèle (1) dispose de suffisamment d'observations pour assurer une bon ajustement de la queue de la distribution F_t. Par ailleurs, τ doit être aussi choisi assez grand de façon à éviter un biais d'estimation due à un mauvais ajustement de la queue de distribution. Nous proposons d'estimer le paramètre τ en utilisant une procédure séquentielle de tests d'adéquations similaire à celle proposée par Grama et Spokoiny (2007-2008) et Durrieu et al. (2012). Dans un premier temps, nous testons $\mathcal{H}_0(\tau)$ l'hypothèse nulle stipulant que F_t est défini par (1) et $s_1, ..., s_m$ une suite d'instants triés par ordre décroissant de sorte que $s_1 \geq ... \geq s_m$ avec m fixé. Dans notre cas, on choisit comme suite s_k les statistiques d'ordre dans la fenêtre de largeur h autour du point t. Nous considérons une suite de tests d'adéquation en déterminant le premier instant s_k notée $\widehat{s}$ pour lequel $\mathcal{H}_0(s_k)$ est rejetée en faveur de l'hypothèse alternative $\mathcal{H}_1(\tau)$: "$F_{t,\tau}$ est la distribution de Pareto avec un point de rupture" où $F_{t,\tau}$ est la fonction de répartition d'excès de F_t au dessus du seuil τ.

Ainsi par cette procédure, nous sélectionnons le meilleur modèle en maximisant par rapport à τ la fonction de vraisemblance pénalisée donnée par :

$$\mathcal{L}_{\tau,h}\left(\tau, \widehat{\theta}_{t,h,\tau}\right) - \mathrm{Pen}_{\tau,h}\left(\tau, \widehat{\theta}_{t,h,\widehat{s}}\right) \qquad \text{où} \qquad \mathrm{Pen}_{t,h}(\tau,\theta) = \mathcal{L}_{t,h}(\tau,\theta),$$

et

$$\mathcal{L}_{t,h}(\tau,\theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}).$$

Nous notons $\widehat{\tau}_{n,t}$ le seuil ainsi obtenu. Le choix du paramètre h est aussi un point crucial. Nous utilisons comme critère la maximisation de la vraisemblance pénalisée calculée aux estimateurs $\widehat{\tau}_{n,t}$ et $\widehat{\theta}_{t,h,\widehat{s}}$:

$$\widehat{h}_n = \arg\max_{h \in \mathcal{H}} \left(\mathcal{L}_{\widehat{\tau}_{n,t},h}\left(\widehat{\tau}_{n,t}, \widehat{\theta}_{t,h,\widehat{\tau}_{n,t}}\right) - \mathrm{Pen}_{\widehat{\tau}_{n,t},h}\left(\widehat{\tau}_{n,t}, \widehat{\theta}_{t,h,\widehat{s}}\right)\right),$$

où $\mathcal{H} = \{h_i : h_i = h_0 q^i, i = 1, \ldots, M\}$ avec $q > 1$, $h_0 > 0$ et M grand.

2 Propriétés asymptotiques

Soit τ_n et h_n deux suites qui réalisent l'équilibre en convergence entre le carré du biais $\mathcal{K}\left(F_{t,\tau}, G_{\tau_n,\theta_{t,\tau_n}}\right)$ et la variance de l'erreur stochastique du modèle calculée sur la fenêtre de taille h_n à l'instant t

$$\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))},$$

où $n_{t,h_n} = \sum_{i=1}^{n} W_{t,h_n}(t_i)$. Nous notons θ_{t,τ_n} le paramètre dit d'oracle.

Théorème 2.1 *Sous des conditions de régularités, nous avons quand $n \to \infty$*

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_{n,t}}, \theta_{\tau_n}\right) = O_P\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right). \tag{4}$$

On en déduit du Théorème 2.1 que

$$\mathcal{K}\left(F_t, \widehat{F}_{t,h_n,\widehat{\tau}_{n,t}}\right) = O_P\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right),$$

quand $n \to \infty$. Nous montrons que la vitesse de convergence obtenue est quasi-optimale.

Dans le cas du modèle de Hall (Hall 1982) défini par

$$F_t(x) = 1 - Cx^{-1/\gamma(t)} - (1-C)x^{-1/\delta(t)} \qquad \text{avec} \qquad x \geq 1 \text{ et } 0 \leq t \leq 1, \tag{5}$$

où $0 < C < 1$, $\gamma(t) > \delta(t)$ sont des fonctions β-Höldériennes avec $0 < \beta \leq 1$, la vitesse de convergence est

$$\mathcal{K}\left(F_t, \widehat{F}_{t,h_n,\widehat{\tau}_{n,t}}\right) = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2\beta}{1+\beta\left(2+\frac{\delta(t)}{\gamma(t)-\delta(t)}\right)}}\right).$$

Ce modèle englobe une grande variété de modèles comme les lois stables (non normales) et la distribution de Fréchet. Si $\delta(t) = 0$, le modèle (5) correspond au modèle de Pareto et la vitesse devient $\left(\frac{\log n}{n}\right)^{\frac{2\beta}{1+2\beta}}$ qui correspond à la vitesse de convergence d'un estimateur d'une fonction non paramétrique du paramètre de régularité β.

3 Etude par simulation

Les propriétés de la méthode proposée sont étudiées sur des simulation en utilisant le modèle (5). Nous fixons pour nos simulations un échantillon de taille $n = 50000$ avec $C = 0.75$, $m = 50$, $M = 100$ et

$$\gamma(t) = 0.5 + 0.25\sin(2\pi t) \qquad \text{et} \qquad \frac{1}{\delta(t)} = \frac{1}{\gamma(t)} + 5.$$

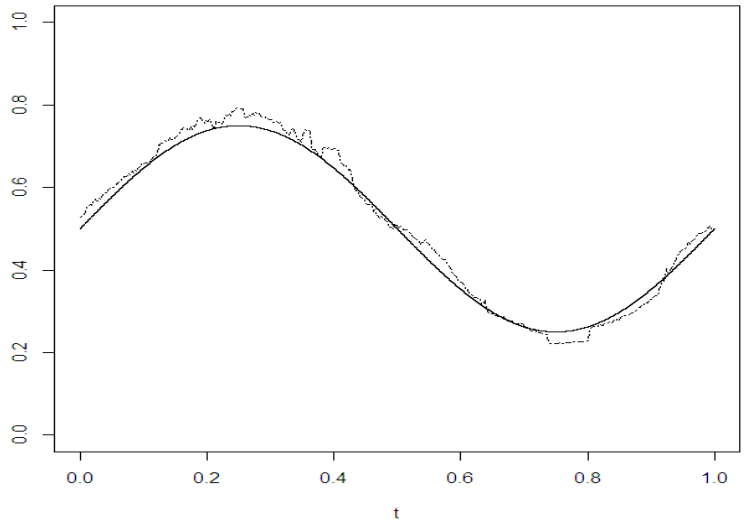

FIGURE 1 – Représentation de l'estimateur adaptatif de $\gamma(t)$ en trait pointillé et de la fonction théorique en trait plein.

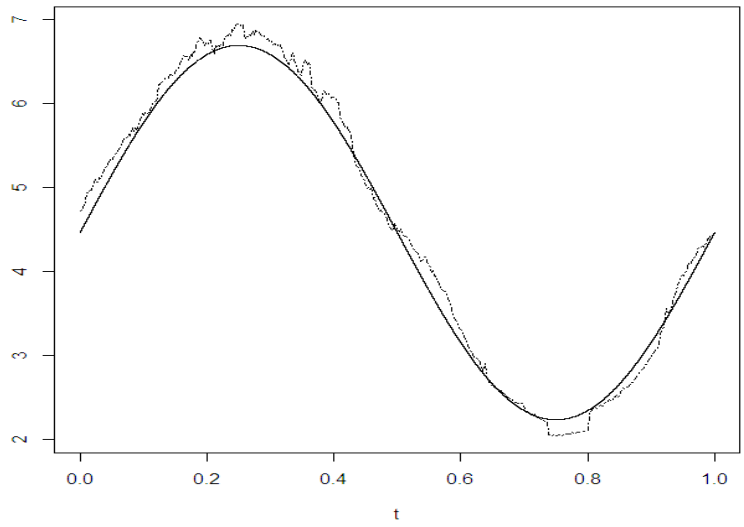

FIGURE 2 – Représentation de l'estimateur adaptatif de $\widehat{q}_{0.9999}(t)$ en trait pointillé et de la fonction théorique en trait plein.

Dans les Figure 1 et 2, nous observons un très bon ajustement de l'estimateur adaptatif du paramètre $\gamma(t)$ et de l'estimateur $\widehat{q}_{0.9999}(t)$. Des analyses similaires effectuées pour les modèles de Pareto avec point de rupture et Fréchet sur un nombre important d'échantillons donnent aussi des résultats satisfaisants. Nous donnerons aussi un exemple d'application de cette procédure sur des données réelles dans le but d'estimer des changements globaux (pollution, changement de température) et ainsi d'aider à la surveillance de systèmes aquatiques (Coudret et al. 2013 ; Schmitt et al. 2011 ; Sow et al. 2011 ; Tran et al. 2011).

Bibliographie

[1] Coudret R., Durrieu. G., Saracco J. (2013) Comparaison of kernel density estimators with assumption on number of modes : application on environmental monitoring data. Communication in statistics, sous presse.

[2] Durrieu G., Grama I., Le Tilly V., Massabuau J.C., Pham Q.K. (2012), Évènements rares sur des séries temporelles environnementales. Proc. de la société Française de Statistique, 6 pages.

[3] Grama I. and Spokoiny V. (2007). Pareto approximation of the tail by local exponential modeling. Bulletin of Academy of Science of Moldova, 53(1), 1-22.

[4] Grama I. and Spokoiny V. (2008) Statistics of extremes by oracle estimation, *Annals of Statistics*, 36(4), 1619-1648.

[5] Hall P. (1982) On some simple estimates of an exponent of regular variation. Journal of the Royal Statistical Society Series B, 44, 37-42.

[6] Schmitt F., De Rosa M., Durrieu G., Sow M., Ciret P., Tran D. and Massabuau, J.C. (2011), Statistical study of bivalve high frequency microclosing behavior : scaling properties and shot noise analysis, *International Journal of Bifurcation and Chaos*, 21(12), 3565-3576.

[7] Schwartzmann C., Durrieu G., Sow M., Ciret P., Lazareth C. and Massabuau J.C. (2011) In situ giant clam growth rate behavior in relation to temperature, *Limnology and Oceanography*, 56(5), 1940-1951.

[8] Sow M., Durrieu G., Briollais L., Ciret P. et Massabuau J. C. (2011) Water quality assessment by means of HFNI valvometry and high-frequency data, *Environmental Modeling and Assessment*, 182, 155-170.

[9] Tran D., Nadau A., Durrieu G., Ciret P., Parisot J.C. and Massabuau J.C. (2011) Field Chronobiology of a Molluscan Bivalve : How the Moon and Sun Cycles Interact to Drive Oyster Activity Rhythms, *Chronobiology International*, 28(4), 307-317.